ITCS 1880: Final Project

By Michael Beebe

Loading in dependencies

library(tidyverse)  
library(glue)  
library(ggplot2)  
library(hrbrthemes)

## Warning: package 'hrbrthemes' was built under R version 4.0.3

Loading in dataset to be analyzed

customer\_data <- read.csv("Data/Mall\_Customers.csv")

# Data Analysis

The number of rows and the number of columns in the data set.

nrow(customer\_data)

## [1] 200

ncol(customer\_data)

## [1] 5

The names of the columns in the data set

colnames(customer\_data)

## [1] "CustomerID" "Gender" "Age"   
## [4] "Annual.Income..k.." "Spending.Score..1.100."

The top 10 rows of the data set

head(customer\_data, 10)

## CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100.  
## 1 1 Male 19 15 39  
## 2 2 Male 21 15 81  
## 3 3 Female 20 16 6  
## 4 4 Female 23 16 77  
## 5 5 Female 31 17 40  
## 6 6 Female 22 17 76  
## 7 7 Female 35 18 6  
## 8 8 Female 23 18 94  
## 9 9 Male 64 19 3  
## 10 10 Female 30 19 72

The class of the dataset and each column

class(customer\_data)

## [1] "data.frame"

lapply(customer\_data, class)

## $CustomerID  
## [1] "integer"  
##   
## $Gender  
## [1] "character"  
##   
## $Age  
## [1] "integer"  
##   
## $Annual.Income..k..  
## [1] "integer"  
##   
## $Spending.Score..1.100.  
## [1] "integer"

Summarize the data in the Mall Customers data set. Be sure to include: Min, Max, Mean, Median, Quartiles and Standard Deviation:

### Age

Min

glue("Min Age: {min(customer\_data$Age)}")

## Min Age: 18

Max

glue("Max Age: {max(customer\_data$Age)}")

## Max Age: 70

Mean

glue("Mean Age: {mean(customer\_data$Age)}")

## Mean Age: 38.85

Median

glue("Median Age: {median(customer\_data$Age)}")

## Median Age: 36

Quartiles

quantile(customer\_data$Age)

## 0% 25% 50% 75% 100%   
## 18.00 28.75 36.00 49.00 70.00

Standard Deviation

glue("Standard Deviation of Age: {sd(customer\_data$Age)}")

## Standard Deviation of Age: 13.9690073315589

### Annual Income ($1,000’s)

Min

glue("Min Annual Income in $1,000's: {min(customer\_data$Annual.Income..k..)}")

## Min Annual Income in $1,000's: 15

Max

glue("Max Annual Income in $1,000's: {max(customer\_data$Annual.Income..k..)}")

## Max Annual Income in $1,000's: 137

Mean

glue("Mean Annual Income in $1,000's: {mean(customer\_data$Annual.Income..k..)}")

## Mean Annual Income in $1,000's: 60.56

Median

glue("Median Annual Income in $1,000's: {median(customer\_data$Annual.Income..k..)}")

## Median Annual Income in $1,000's: 61.5

Quartiles

quantile(customer\_data$Annual.Income..k..)

## 0% 25% 50% 75% 100%   
## 15.0 41.5 61.5 78.0 137.0

Standard Deviation

glue("Standard Deviation of Annual Income in $1,000's: {sd(customer\_data$Annual.Income..k..)}")

## Standard Deviation of Annual Income in $1,000's: 26.2647211652712

### Spending Score (1-100)

Min

glue("Min Spending Score: {min(customer\_data$Spending.Score..1.100.)}")

## Min Spending Score: 1

Max

glue("Max Spendinc Score: {max(customer\_data$Spending.Score..1.100.)}")

## Max Spendinc Score: 99

Mean

glue("Mean Spending Score: {mean(customer\_data$Spending.Score..1.100.)}")

## Mean Spending Score: 50.2

Median

glue("Median Spending Score: {median(customer\_data$Spending.Score..1.100.)}")

## Median Spending Score: 50

Quartiles

quantile(customer\_data$Spending.Score..1.100.)

## 0% 25% 50% 75% 100%   
## 1.00 34.75 50.00 73.00 99.00

Standard Deviation

glue("Standard Deviation of Spending Score: {sd(customer\_data$Spending.Score..1.100.)}")

## Standard Deviation of Spending Score: 25.8235216683702

### Observations by Gender

Average age by gender

age\_by\_gender <- aggregate(  
 x = customer\_data$Age,  
 by = list(customer\_data$Gender),  
 FUN = mean  
)  
age\_by\_gender

## Group.1 x  
## 1 Female 38.09821  
## 2 Male 39.80682

Average annual income by gender

income\_by\_gender <- aggregate(  
 x = customer\_data$Annual.Income..k..,  
 by = list(customer\_data$Gender),  
 FUN = mean  
)  
income\_by\_gender

## Group.1 x  
## 1 Female 59.25000  
## 2 Male 62.22727

Average spending score by gender

spending\_by\_gender <- aggregate(  
 x = customer\_data$Spending.Score..1.100.,  
 by = list(customer\_data$Gender),  
 FUN = mean  
)  
spending\_by\_gender

## Group.1 x  
## 1 Female 51.52679  
## 2 Male 48.51136

New DF combining the three above aggregations

summarized\_avgs\_by\_gender <- data.frame(  
 Gender = c("Female", "Male"),  
 Average\_Age = age\_by\_gender$x,  
 Average\_Annual\_Income = income\_by\_gender$x,  
 Avergae\_Spending = spending\_by\_gender$x  
)  
summarized\_avgs\_by\_gender

## Gender Average\_Age Average\_Annual\_Income Avergae\_Spending  
## 1 Female 38.09821 59.25000 51.52679  
## 2 Male 39.80682 62.22727 48.51136

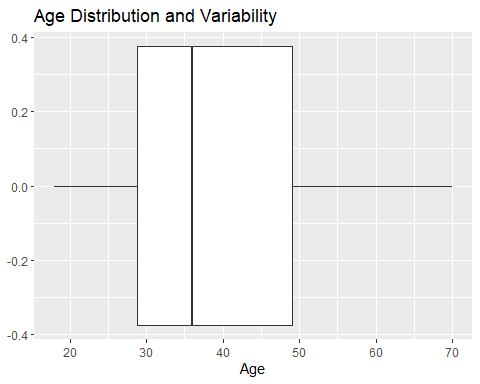
### Data Analysis Summary

!!!!!!!! TODO !!!!!!!!

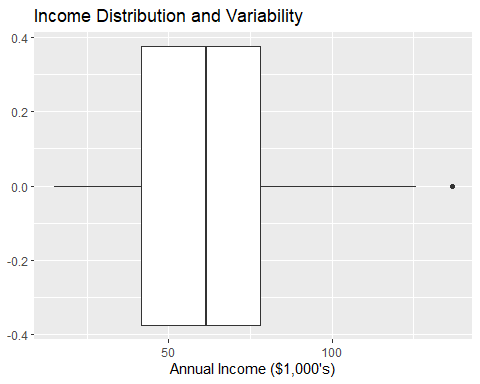
# Plotting

### Boxplots

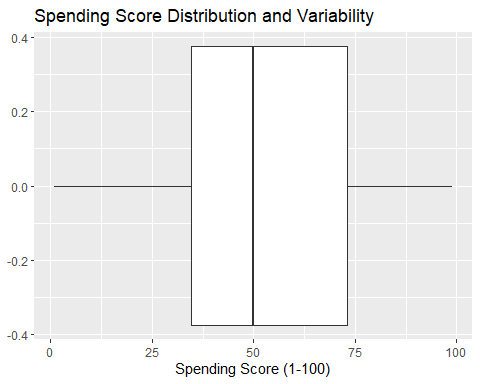
ggplot(data = customer\_data, aes(x = Age)) +  
 geom\_boxplot() +  
 ggtitle("Age Distribution and Variability") +  
 xlab("Age") +  
 theme\_get()



ggplot(data = customer\_data, aes(x = Annual.Income..k..)) +  
 geom\_boxplot() +  
 ggtitle("Income Distribution and Variability") +  
 xlab("Annual Income ($1,000's)") +  
 theme\_get()

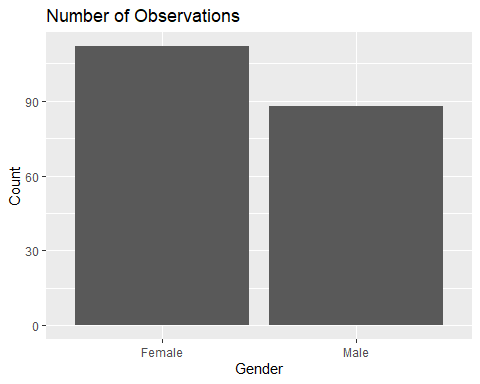


ggplot(data = customer\_data, aes(x = Spending.Score..1.100.)) +  
 geom\_boxplot() +  
 ggtitle("Spending Score Distribution and Variability") +  
 xlab("Spending Score (1-100)") +  
 theme\_get()



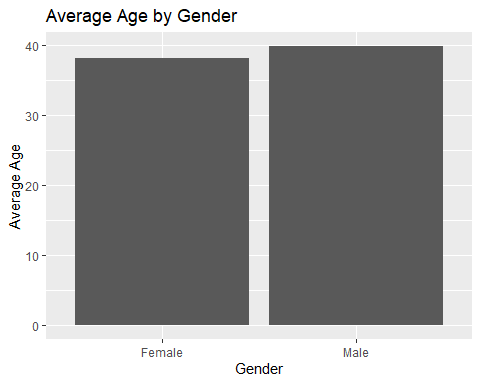
### Histograms (By Gender)

ggplot(data = customer\_data, mapping = aes(x=Gender)) +  
 geom\_bar() +  
 ggtitle ("Number of Observations") +  
 xlab("Gender") +  
 ylab("Count") +  
 theme\_get()



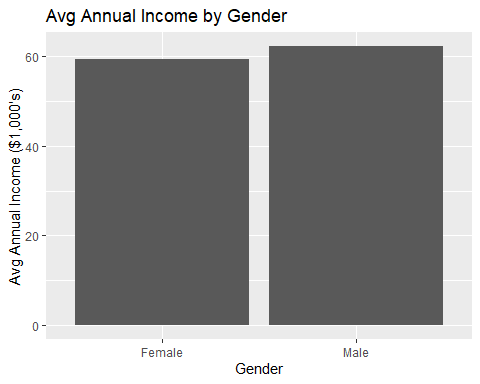
ggplot(data = summarized\_avgs\_by\_gender, mapping = aes(x=Gender, y=Average\_Age)) +  
 geom\_histogram(stat="identity") +  
 ggtitle ("Average Age by Gender") +  
 xlab("Gender") +  
 ylab("Average Age") +  
 theme\_get()

## Warning: Ignoring unknown parameters: binwidth, bins, pad



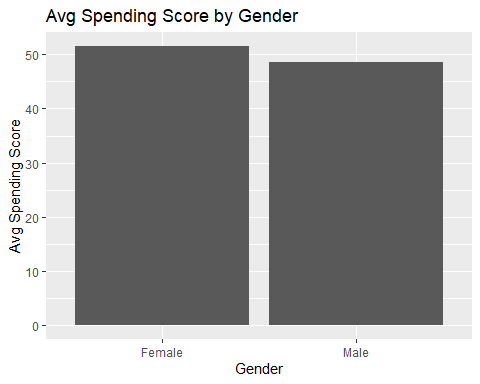
ggplot(data = summarized\_avgs\_by\_gender, mapping = aes(x=Gender, y=Average\_Annual\_Income)) +  
 geom\_histogram(stat="identity") +  
 ggtitle ("Avg Annual Income by Gender") +  
 xlab("Gender") +  
 ylab("Avg Annual Income ($1,000's)") +  
 theme\_get()

## Warning: Ignoring unknown parameters: binwidth, bins, pad



ggplot(data = summarized\_avgs\_by\_gender, mapping = aes(x=Gender, y=Avergae\_Spending)) +  
 geom\_histogram(stat="identity") +  
 ggtitle ("Avg Spending Score by Gender") +  
 xlab("Gender") +  
 ylab("Avg Spending Score") +  
 theme\_get()

## Warning: Ignoring unknown parameters: binwidth, bins, pad

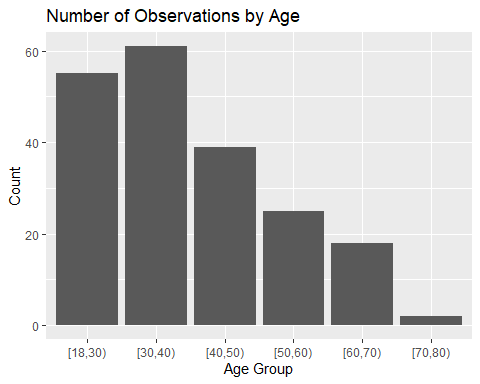


### Histograms (By Age Group)

# ADDING AN AGE GROUP COLUMN TO THE CUSTOMER DATA DF  
age\_cut <- cut(customer\_data$Age, breaks = c(18, 30, 40, 50, 60, 70, 80), right = FALSE)  
customer\_data\_2 <- data.frame(customer\_data, Age\_Group = age\_cut)  
head(customer\_data\_2)

## CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100. Age\_Group  
## 1 1 Male 19 15 39 [18,30)  
## 2 2 Male 21 15 81 [18,30)  
## 3 3 Female 20 16 6 [18,30)  
## 4 4 Female 23 16 77 [18,30)  
## 5 5 Female 31 17 40 [30,40)  
## 6 6 Female 22 17 76 [18,30)

ggplot(customer\_data\_2, aes(x = Age\_Group)) +  
 geom\_bar() +  
 ggtitle ("Number of Observations by Age") +  
 xlab("Age Group") +  
 ylab("Count") +  
 theme\_get()



income\_by\_age\_group <- aggregate(  
 x = customer\_data\_2$Annual.Income..k..,  
 by = list(customer\_data\_2$Age\_Group),  
 FUN = mean  
)  
income\_by\_age\_group

## Group.1 x  
## 1 [18,30) 52.54545  
## 2 [30,40) 70.18033  
## 3 [40,50) 63.94872  
## 4 [50,60) 58.68000  
## 5 [60,70) 49.16667  
## 6 [70,80) 47.50000

spending\_by\_age\_group <- aggregate(  
 x = customer\_data\_2$Spending.Score..1.100.,  
 by = list(customer\_data\_2$Age\_Group),  
 FUN = mean  
)  
spending\_by\_age\_group

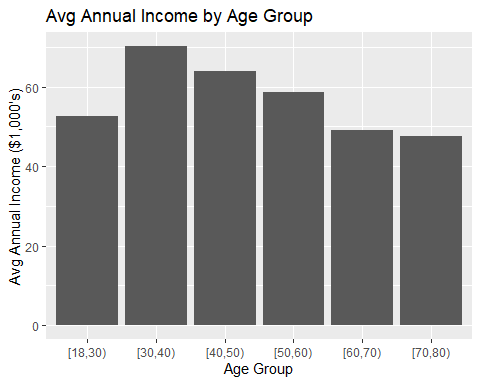
## Group.1 x  
## 1 [18,30) 58.58182  
## 2 [30,40) 61.09836  
## 3 [40,50) 34.94872  
## 4 [50,60) 34.72000  
## 5 [60,70) 41.61111  
## 6 [70,80) 55.50000

summarized\_avgs\_by\_age\_group <- data.frame(  
 Age\_Group = income\_by\_age\_group$Group.1,  
 Avg\_Income = income\_by\_age\_group$x,  
 Avg\_Spending = spending\_by\_age\_group$x  
)  
summarized\_avgs\_by\_age\_group

## Age\_Group Avg\_Income Avg\_Spending  
## 1 [18,30) 52.54545 58.58182  
## 2 [30,40) 70.18033 61.09836  
## 3 [40,50) 63.94872 34.94872  
## 4 [50,60) 58.68000 34.72000  
## 5 [60,70) 49.16667 41.61111  
## 6 [70,80) 47.50000 55.50000

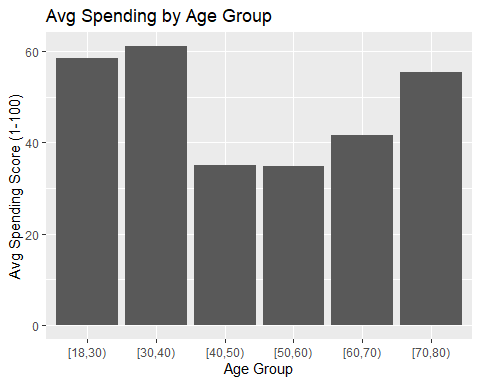
ggplot(data = summarized\_avgs\_by\_age\_group, mapping = aes(x=Age\_Group, y=Avg\_Income)) +  
 geom\_histogram(stat="identity") +  
 ggtitle ("Avg Annual Income by Age Group") +  
 xlab("Age Group") +  
 ylab("Avg Annual Income ($1,000's)") +  
 theme\_get()

## Warning: Ignoring unknown parameters: binwidth, bins, pad



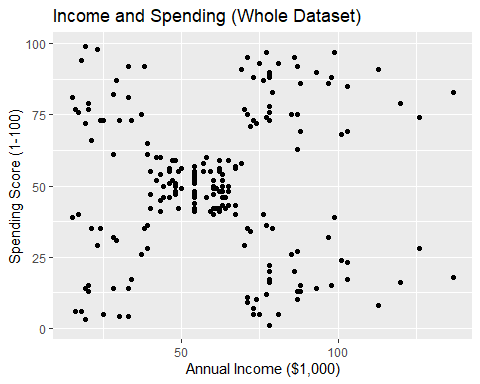
ggplot(data = summarized\_avgs\_by\_age\_group, mapping = aes(x=Age\_Group, y=Avg\_Spending)) +  
 geom\_histogram(stat="identity") +  
 ggtitle ("Avg Spending by Age Group") +  
 xlab("Age Group") +  
 ylab("Avg Spending Score (1-100)") +  
 theme\_get()

## Warning: Ignoring unknown parameters: binwidth, bins, pad



### Scatterplot

ggplot(data = customer\_data, mapping = aes(x=Annual.Income..k.., y=Spending.Score..1.100.)) +  
 geom\_point() +  
 ggtitle("Income and Spending (Whole Dataset)") +  
 xlab("Annual Income ($1,000)") +  
 ylab("Spending Score (1-100)") +  
 theme\_get()



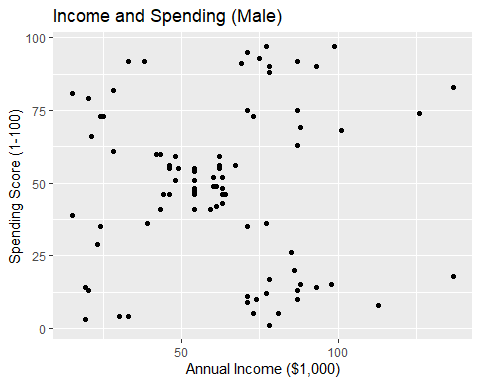
male\_df <- customer\_data\_2[customer\_data\_2$Gender == "Male", ]  
head(male\_df)

## CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100. Age\_Group  
## 1 1 Male 19 15 39 [18,30)  
## 2 2 Male 21 15 81 [18,30)  
## 9 9 Male 64 19 3 [60,70)  
## 11 11 Male 67 19 14 [60,70)  
## 15 15 Male 37 20 13 [30,40)  
## 16 16 Male 22 20 79 [18,30)

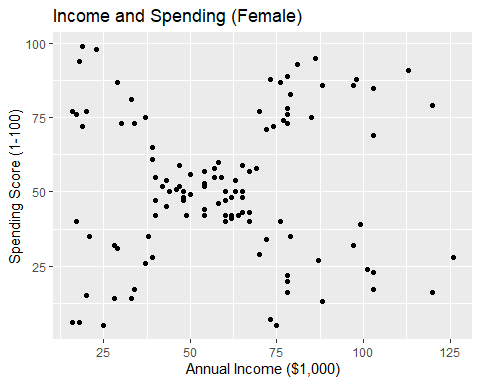
female\_df <- customer\_data\_2[customer\_data\_2$Gender == "Female", ]  
head(female\_df)

## CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100. Age\_Group  
## 3 3 Female 20 16 6 [18,30)  
## 4 4 Female 23 16 77 [18,30)  
## 5 5 Female 31 17 40 [30,40)  
## 6 6 Female 22 17 76 [18,30)  
## 7 7 Female 35 18 6 [30,40)  
## 8 8 Female 23 18 94 [18,30)

ggplot(data = male\_df, mapping = aes(x=Annual.Income..k.., y=Spending.Score..1.100.)) +  
 geom\_point() +  
 ggtitle("Income and Spending (Male)") +  
 xlab("Annual Income ($1,000)") +  
 ylab("Spending Score (1-100)") +  
 theme\_get()



ggplot(data = female\_df, mapping = aes(x=Annual.Income..k.., y=Spending.Score..1.100.)) +  
 geom\_point() +  
 ggtitle("Income and Spending (Female)") +  
 xlab("Annual Income ($1,000)") +  
 ylab("Spending Score (1-100)") +  
 theme\_get()



### Plot Analysis Summary

!!!!!!!! TODO !!!!!!!!

# Probability and Statistical Analysis